Journal of Biomedical Informatics 45 (2012) 528–534

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

MysiRNA: Improving siRNA efficacy prediction using a machine-learning model combining multi-tools and whole stacking energy (ΔG)

Mohamed Mysara [a,b,c,*], Mahmoud Elhefnawi [a,b,*], Jonathan M. Garibaldi [c,*]

[a] Informatics and Systems Department and Biomedical Informatics and Chemoinformatics Group, Division of Engineering Research and Centre of Excellence for Advanced Sciences, National Research Centre, Tahrir Street, 12311 Cairo, Egypt

[b] Information Technology Institute, Division of Biomedical Informatics and Bioinformatics, Smart Village, 6th of October City, Egypt

[c] Intelligent Modelling and Analysis Research Group, School of Computer Science, University of Nottingham, Jubilee Campus, Wollaton Road, Nottingham NG8 1BB, UK

ARTICLE INFO

Article history:
Received 19 April 2011
Accepted 15 February 2012
Available online 25 February 2012

Keywords:
siRNA efficiency prediction
siRNA functionality prediction
Artificial neural network
Whole stacking energy
Gibbs energy
ΔG

ABSTRACT

The investigation of small interfering RNA (siRNA) and its posttranscriptional gene-regulation has become an extremely important research topic, both for fundamental reasons and for potential longer-term therapeutic benefits. Several factors affect the functionality of siRNA including positional preferences, target accessibility and other thermodynamic features. State of the art tools aim to optimize the selection of target siRNAs by identifying those that may have high experimental inhibition. Such tools implement artificial neural network models as *Biopredsi* and *ThermoComposition21*, and linear regression models as *DSIR*, *i-Score* and *Scales*, among others. However, all these models have limitations in performance. In this work, a neural-network trained new siRNA scoring/efficacy prediction model was developed based on combining two existing scoring algorithms (*ThermoComposition21* and *i-Score*), together with the whole stacking energy (ΔG), in a multi-layer artificial neural network. These three parameters were chosen after a comparative combinatorial study between five well known tools. Our developed model, '*MysiRNA*' was trained on 2431 siRNA records and tested using three further datasets. *MysiRNA* was compared with 11 alternative existing scoring tools in an evaluation study to assess the predicted and experimental siRNA efficiency where it achieved the highest performance both in terms of correlation coefficient ($R^2 = 0.600$) and receiver operating characteristics analysis (AUC = 0.808), improving the prediction accuracy by up to 18% with respect to sensitivity and specificity of the best available tools. *MysiRNA* is a novel, freely accessible model capable of predicting siRNA inhibition efficiency with improved specificity and sensitivity. This multiclassifier approach could help improve the performance of prediction in several bioinformatics areas. *MysiRNA model*, part of *MysiRNA*-Designer package [1], is expected to play a key role in siRNA selection and evaluation.

1. Introduction

Small interfering RNAs (siRNAs) are one of the cellular main processes for posttranscriptional gene modification, capable of down regulating mRNA expression and causing targeted gene silencing. These small RNA molecules are one of the cellular defense mechanisms that act not only against exogenous genetic material (such as viruses) but also against endogenous genes [2]. Due to their small length (19–21 nucleotides), they can bypass the interferon-response responsible for undesirable cell death, and hence are more capable of producing a desired silencing action [3]. This induced gene silencing may be used to identify gene functions or even (as an ultimate goal) to treat several gene-mediated diseases such as cancer, AIDS and neurodegenerative disorders for which no proper cure has yet been found [4–8].

In the last decade, siRNAs have become a major interest of many biologists due to their selectivity and potency. This has resulted in their therapeutic application for several viral-mediated diseases such as Influenza A virus, HIV, Hepatitis B virus, and RSV viruses and in cancer clinical trials [9–12]. As a result, siRNA silencing is considered one of the most promising techniques in future therapy, and predicting their inhibition efficiency is crucial for proper siRNA selection. As a targeted gene could have thousands of potential siRNAs, finding the most active siRNA among them constitutes a huge challenge facing researchers in this field.

Several algorithms have been developed to predict siRNA activity. However, only a few of them have achieved an acceptable level of specificity and sensitivity. These algorithms could be

* Corresponding authors. Address: Biomedical Informatics and Chemoinformatics Group, National Research Centre, Tahrir Street, 12311 Cairo, Egypt. Fax: +2 02 33370931 (M. Mysara, M. Elhefnawi).

E-mail addresses: mohamed.mysara@nottingham.ac.uk, mm.abdelwahab@nrc.sci.eg (M. Mysara), mahef@aucegypt.edu (M. Elhefnawi), jmg@cs.nott.ac.uk (J.M. Garibaldi).

1532-0464/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2012.02.005

subclassified into two groups, namely 'first generation' and 'second generation' tools. The first generation tools select the most efficient siRNAs based on differential ends thermodynamic stability measures, mRNA secondary structure and base preferences specific position target uniqueness; among these are Reynolds et al. [13], Amarzguioui and Prydz [14], Takasaki et al. [15], Katoh and Suzuki [16], Ui-Tei et al. [17], and Hsieh et al. [18]. However, these first generation tools were shown to have low accuracy, as up to 65% of the siRNAs predicted as active failed to achieve 90% inhibition when tested experimentally and up to 20% of them were found to be inactive [19]. Consequently, it was perceived that there was a need to develop techniques that incorporated machine-learning to interpret and utilize the experimentally obtained data.

The second generation tools were developed by applying sophisticated data mining techniques to interpret annotated records of siRNA with their experimental inhibition. It was not until the introduction of the 'Huesken' dataset, used to train *Biopredsi's* artificial neural network model, that a breakthrough in the prediction accuracy was attained [20]. Further improvement in prediction accuracy was brought about by the use of a simplified linear regression model in '*DSIR*' [21]. *ThermoComposition21* [22] combined position dependent features together with thermodynamic features in one artificial neural network model, hence, improving the prediction accuracy. Two more models (*i-Score* and *Scales* [23,24]) were developed that used simpler linear regression models to achieve similar levels of performance.

These second generation models (whether artificial neural network or linear regression based) perform significantly better than the first generation tools. However, the prediction accuracy and statistical significance of the five mentioned algorithms were found to be very similar to each other when they were tested against a new dataset that had not been used in their training as reported in [23,24].

We investigated the top five scoring algorithms (*Biopredsi*, *DSIR*, *ThermoComposition21*, *i-Score* and *Scales*), aiming to build a model that combines their different scoring algorithms and features in a single model in order to improve the prediction accuracy. In addition, we studied the effect of including whole stacking energy (ΔG) as an independent parameter for enhancing the specificity and sensitivity of siRNA efficiency prediction. Receiver operating characteristics (ROCs) curve analysis, Matthews correlation coefficient (MCC) and Pearson's correlation coefficient (R^2) were used for the validation approach. This model was implemented in the MysiRNA-Designer workflow package for selection and design of efficient siRNAs .

2. Material and methods

2.1. Dataset selection

In this work, the Huesken dataset (dataset A) was used to train our model, the quality of this dataset is ensured by the Gaussian distribution of their potencies [20,23]. Dataset A consists of 2431 siRNAs with their experimental inhibition efficiency, and has previously been used to train *Biopredsi*, *DSIR*, *ThermoComposition21* and *i-Score*. Datasets B and C were used for validation step, where dataset B consisted of 419 siRNA records from five different publications, Reynolds (60%), Vickers (18%), Haborth (11%), Ui-Tie (9%) and Khovorova (2%). These datasets had been collected together and published in the work of Ichihara in the development and evaluation of the *i-Score* software [23]. There are two versions of ThermoComposition, one for designing siRNA with length 19 mer, other with 21 mer. Some records from dataset B were used to train *ThermoComposition19* [22]. We filtered dataset B and found 38 siRNA unique records that were not involved in training of any of the second generation tools, and these 38 records were used to create dataset C. Scoring of the siRNA datasets was performed using *i-Score* designer that includes several first and second generation algorithms. To include the *Scales* prediction tool in our evaluation, we calculated and added its predicted scores to each of the datasets A, B and C.

Another data was used consisting of nine genes with 18,593 siRNA presented by Fellmann et al. [29], in which each possible siRNAs were designed and experimentally tested. The ratio between active and inactive siRNA was found to be (1:77), to overcome the data being skewed to negative results, the positive incidences were isolated (238) and randomly selected data from the remaining, negative, data was isolated to form a total of 476 siRNA records, named dataset D (see Table 1). All these datasets are available through supplementary file, named dataset A, B, C and D, respectively.

2.2. Parameter selection

We investigated several second generation tools in an attempt to build a model that combines several scoring techniques. We included five models (Biopredsi, *DSIR*, *ThermoComposition21*, *i-Score* and *Scales*) in our study and implemented each of these scores as a parameter in our model. As in Ichihara's work, we used the simulated Biopredsi (*s-Biopredsi* [23]) rather than the original Biopredsi [20] due to the inaccessibility of the original model (the correspondence between s-Biopredsi and Biopredsi was confirmed by achieving a Pearson correlation coefficient of 1 and by achieving identical receiver operating characteristics ROC analysis). We considered all possible combinations of models to determine whether any provided an appreciable improvement in the prediction accuracy. After identifying that the best combination consisted of two of the scores, we added a third parameter to our model, the whole stacking energy (ΔG), evaluating its ability to further improve the accuracy of results.

2.3. Training and validation

Due to the non-linear relationship between the selected parameters, we could not apply linear regression to build our model. Therefore, we built our model using multilayer artificial neural networks implemented in WEKA software version 3.7.1 (Waikato Environment for Knowledge Analysis, University of Waikato, NZ)[1].

As a validation step we used the Pearson correlation coefficient to represent the correlation between the predicted and actual siRNA inhibition efficiency. In addition, we used receiver operating characteristic (ROC) analysis that combines both sensitivity and specificity by plotting the sensitivity (Y axis) against 1-specificity (X axis). It is then possible to calculate the area under the curve, known as the AUC, as a single measure of performance (for which an AUC of 1 reflects perfect classification and an AUC of 0.5 reflects random classification). Moreover, Matthews correlation coefficient (MCC) was also included. MCC is a correlation coefficient between the observed and predicted binary classifications; taking into account false positive (FP), false negative (FN), true positive (TP) and true negative (TN), it returns a value between −1 (negative correlation), 0 (no correlation) and +1 (positive correlation).

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FB})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \quad (1)$$

[1] The software is a freeware, available online at http://www.cs.waikato.ac.nz/ml/weka/.

Table 1
Comparison between datasets A, B, C and D. Comparison between Dataset A, B, C and D illustrating their source, instance where they were used to for model training and testing models. Further analysis of these data showing their size, number of records with percentage inhibition between (50% and 70%), (70% and 90%), (above 90%) and number of genes involved in each dataset. This analysis illustrate the uniformity of each dataset and illustrate the instance where over fitting of the data might occur.

Dataset Name	Dataset source	Train set for	Test set for	Size	50–70% inhibition	70–90% inhibition	>90% inhibition	Num of genes
Dataset A	Novartis	All of the 2nd generation tools + *MysiRNA*	Biopredsi, *DSIR*	2431	778	853	369	30
Dataset B	Reynold, Vickers, Haborth, Ui-Tie, Khovorova	Thermo Composition19	*i-Score*, *Scales*	419	60	117	96	12
Dataset C*	Records extracted from Dataset B	None	*i-Score*, *Scales*	38	7	12	19	7
Dataset D	Fellmann et al. [29]	None	–	476	70	53	127	9

* Dataset C extracted from Dataset B by filtering siRNA records used from training ThermoComposition19, forming 38 unique siRNA records that has not been used for training any of the second generation tools previously.

3. Results

3.1. Parameters selection

Since the siRNA design and scoring tools use different features and weights in the design process, combining these features by building a model containing these scores was attempted. The predicted scores for each of the five algorithms against database A was used as parameters to build models considering all possible combinations. The combinations consisted of either two, three, four and five tools, resulting in 25 possible combinations in total (Table 2). Both linear regression and artificial neural networks (ANNs) were used to combine each model, and used to perform a comparative analysis between these models and their prediction accuracy against database B and C.

It was found that among these, the model that combines *ThermoComposition21* and *i-Score* produced the most accurate results, irrespective of whether linear combination or non-linear (ANN) combination was used. For each combination, the prediction accuracy evaluated by either Pearson correlation coefficient or AUC achieved the highest score (Fig. 1).

Table 2
Different combinations of tools and their prediction accuracy against dataset B. Illustration of the results all possible combination of different scoring tools their analysis against dataset B using both ROC analysis and Pearson correlation coefficient for evaluation. Where the highest accuracy was achieve by the model combing *ThermoComposition21* and i-Score was later implemented in MysiRNA. (A) s-Biopredsi model, (B) DISR model, (C) *ThermoComposition21*, (D) i-Score model, E: Scales model black cells = no linear correlation. Yellow = best combination.

Combination	Linear Regression		Artificial Neural Network		Combination	Linear Regression		Artificial Neural Network	
	Pearson	ROC	Pearson	ROC		Pearson	ROC	Pearson	ROC
A+B	0.557	0.792	0.556	0.790	A+B+E	0.560	0.794	0.553	0.789
A+C	0.568	0.791	0.558	0.786	A+C+D			0.560	0.787
A+D			0.548	0.782	A+C+E			.560	0.786
A+E	0.544	0.779	0.544	0.780	A+D+E	0.544	0.780	.544	0.781
B+C	0.568	0.799	0.562	0.795	B+C+D	0.565	0.798	0.564	0.793
B+D			0.553	0.792	B+C+E	0.570	0.801	0.568	0.795
B+E			0.547	0.787	B+D+E			0.548	0.785
C+D	0.581	0.796	0.581	0.790	C+D+E	0.566	0.788	0.566	0.783
C+E	0.555	0.781	0.443	0.778	B+C+D+E	0.565	0.798	0.564	0.792
D+E	0.540	0.775	0.539	0.774	A+C+D+E	0.568	0.791	0.560	0.787
A+B+C			0.562	0.794	A+B+D+E			0.546	0.797
A+B+D	0.551	0.790	0.549	0.788	A+B+C+E	0.573	0.800	0.566	0.791
					A+B+C+D	0.562	0.795	0.562	0.791

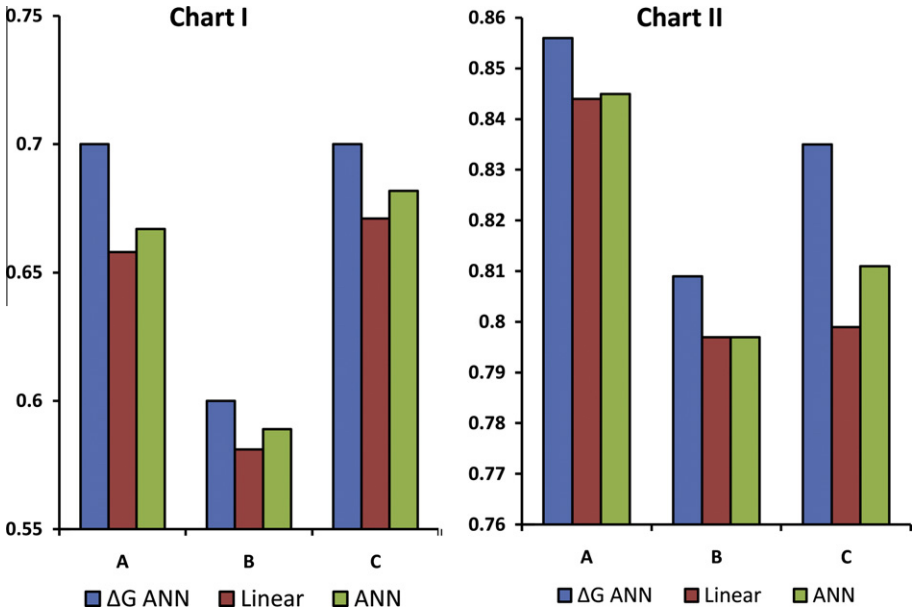

Fig. 1. Comparison between three proposed models, indicating the most accurate model among them. (A) model (blue color) combines *i-Score* + *ThermoComposition21* + whole ΔG in ANN model. (B) Model (Dark red color) combines *i-Score* and *ThermoComposition21* in linear regression model. (C) Model (Green color) combines *i-Score* and *ThermoComposition21* in ANN model. The results were evaluated by both Pearson correlation coefficient (Chart I) and ROC analysis (Chart II) including dataset A, B and C. It was found that ANN model of *ThermoComposition21* + *i-Score* + whole ΔG has the best performance among them. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Inclusion of whole ΔG in the selected model

Whole ΔG was included as an independent parameter to our proposed model. When the three parameters (*ThermoComposition21*, *i-Score* and whole ΔG) were combined using an ANN, a noticeable improvement in the prediction accuracy was obtained. After optimizing the ANN models, this non-linear combination resulted in a Pearson correlation of 0.600 and AUC of 0.808 between the predicted and experimental siRNA efficiency, for dataset B (Fig. 1). We term this model '*MysiRNA*'.

3.3. Comparing MysiRNA with the second generation algorithms

To test the efficiency of the developed model (*MysiRNA*), database B was used in a comparative analysis with other second generation algorithms (Biopredsi, *DSIR*, *ThermoComposition21*, *i-Score* and *Scales*). The actual siRNA activity was plotted against the predicted activity by *MysiRNA* and each of these five techniques, one at a time. The prediction accuracy was verified using Pearson correlation coefficient and AUC. From examination of the Pearson correlation coefficient, *MysiRNA* achieved better correlation with the experimental data against the other techniques (Fig. 2) and brought improvement to the AUC (AUC = 0.808), using 70% as a threshold for siRNA efficiency acceptance (Fig. 3).

3.4. Evaluating MysiRNA model against first and second generation algorithms

Another experiment was then conducted to compare *MysiRNA* with first generation tools including Reynolds, Amarzguioui, Katoh, Hsieh, Takasaki and Ui-Tie, and with second generation tools *s-Biopredsi*, *DSIR*, *ThermoComposition21* and *i-Score*, using all three datasets. The results were then analyzed using Pearson correlation coefficient and AUC. In the ROC analysis, siRNA with inhibition equal to 70% or above is considered active siRNA and below 70% are considered inactive (Fig. 4). As some of the datasets had been used to train certain tools, there was a possibility of over-fitting such that the performance could have been overestimated. Hence, we used Dataset C that is pure/non-biased set. It was found that the *MysiRNA* model provided the most accurate prediction results compared to the 11 models included in this experiment, achieving AUCs of 0.855, 0.808 and 0.834, and Pearson correlations of 0.687, 0.600 and 0.699, against dataset A, B and C, respectively.

To evaluate any benefit of *MysiRNA* over the other models individually, a comparative study was performed between the instances where each of the second-generation models failed to predict siRNA activity, while *MysiRNA* succeeded. It was found that *MysiRNA* increases the prediction capabilities of all of the second generation tools by between 3% and 18% (Table 3).

3.5. Mathew correlation coefficient for models evaluation

Another comparative analysis was performed involving MysiRNA, first generation and second generation models using dataset D, where MCC was used to illustrate the statistical difference between all of the models involved. A threshold of 85% was applied as a cut-off score, above which the prediction is positive and below this threshold siRNA considered inactive. MysiRNA achieved the highest MCC with the experimentally verified data, as it was able to achieve a correlation of 0.51, while the others achieved smaller correlations, see (Table 4). For further details refer to the Supplementary data.

4. Discussion

A number of different models have previously been developed to evaluate siRNA efficiency and predict their inhibition activity. In this work, we studied the best scoring techniques currently available, and their corresponding algorithms, searching for a model that combines the predicted results of these tools to produce more accurate prediction.

We found that combining *ThermoComposition21* with *i-Score* produced the most accurate model among the different combinations of models tested. In addition, we tested the effect of including

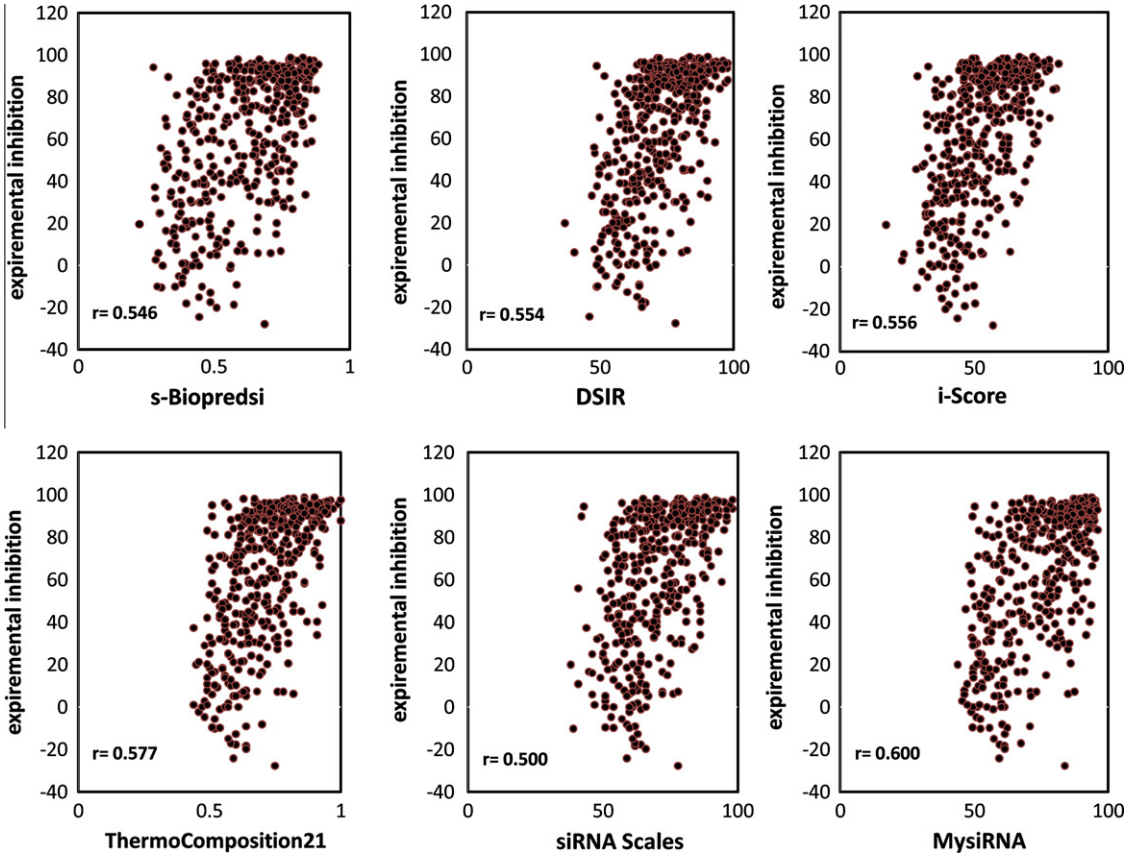

Fig. 2. Comparison between the second generation models and MysiRNA Model using Pearson correlation analysis. Experimental siRNAs activities of dataset B (which has never been used to train any of these scoring models) were plotted against the predicted siRNAs activities by each of the second general scoring tools (*s-Biopredsi*, *DSIR*, *ThermoComposition21*, *i-Score* and Scales) together with MysiRNA model. Pearson correlation coefficient (r) was calculated for each of the six scoring tools. Pearson correlation coefficient of MysiRNA model showed improvement in the performance compared to the other five models. Dataset B was included in this analysis as it was not used for training of any of the six models to avoid over fitting.

the whole stacking energy (ΔG) (which reflects the stability between siRNA duplexes) to our designed model. This resulted in a model with even better performance.

The *Thermocomposition21* algorithm is based on three different parameters, the 'position dependant consensus', the 'dinucleotide content index' and the 'thermodynamic profile'. The position dependent consensus depends on the identification of several desirable and undesirable residues, the dinucleotide content index is based on the number of dinucleotide combinations exceeding the random distribution, and the thermodynamic profile is related to the difference in free energy between 5′ and 3′ [22]. Meanwhile, the *i-Score* algorithm is based on annotated position dependent residues G/C at position (18, 19) and A/T stretch at the 5′ end of the antisense and other parameters that have been previously identified [20,22,25,26]. The fact that these two scores incorporate different features may explain why a combination of the two produces better performance.

Whole siRNA stacking energy (Gibbs energy, ΔG) reflects the stability of siRNA duplex [27]. It depends on the stacking energy difference between each dinucleotide and their complementary nucleotides in the siRNA duplex. It has been shown that there is a correlation between siRNA inhibition efficiency and whole stacking energy [23,28]. While whole ΔG was found to not be of benefit within the i-Score model [23], we suggest that this could be due to the fact that the non-linear correlation between whole ΔG and the other parameters could not be adequately represented within i-Score's linear model. However, the addition of ΔG as a standalone parameter to the MysiRNA model appreciably enhanced the prediction of the model (the Pearson correlation increased from 0.581 to 0.600).

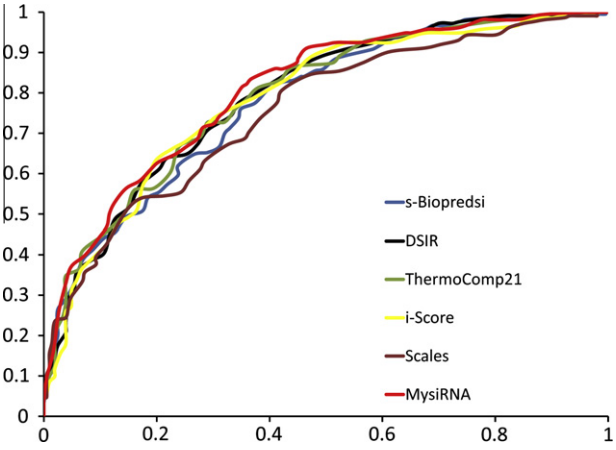

Fig. 3. ROC analysis showing area under the curve of the second generation tools and MysiRNA. Comparative analysis including *s-Biopredsi*, *DSIR*, *ThermoComposition21*, *i-Score*, Scales and MysiRNA to test the sensitivity (Y axes) and 1- specificity (X axes) of the predicted results with of these tools. The level of specificity and sensitivity is slightly improved by using MysiRNA compared to the other tools as MysiRNA scored 0.808 compared to the AUC achieved by s-Biopredsi, *DSIR*, *ThermoComposition21*, *i-Score* and Scales (0.782,0 .792, 0.794, 0.756, 0.781) respectively.

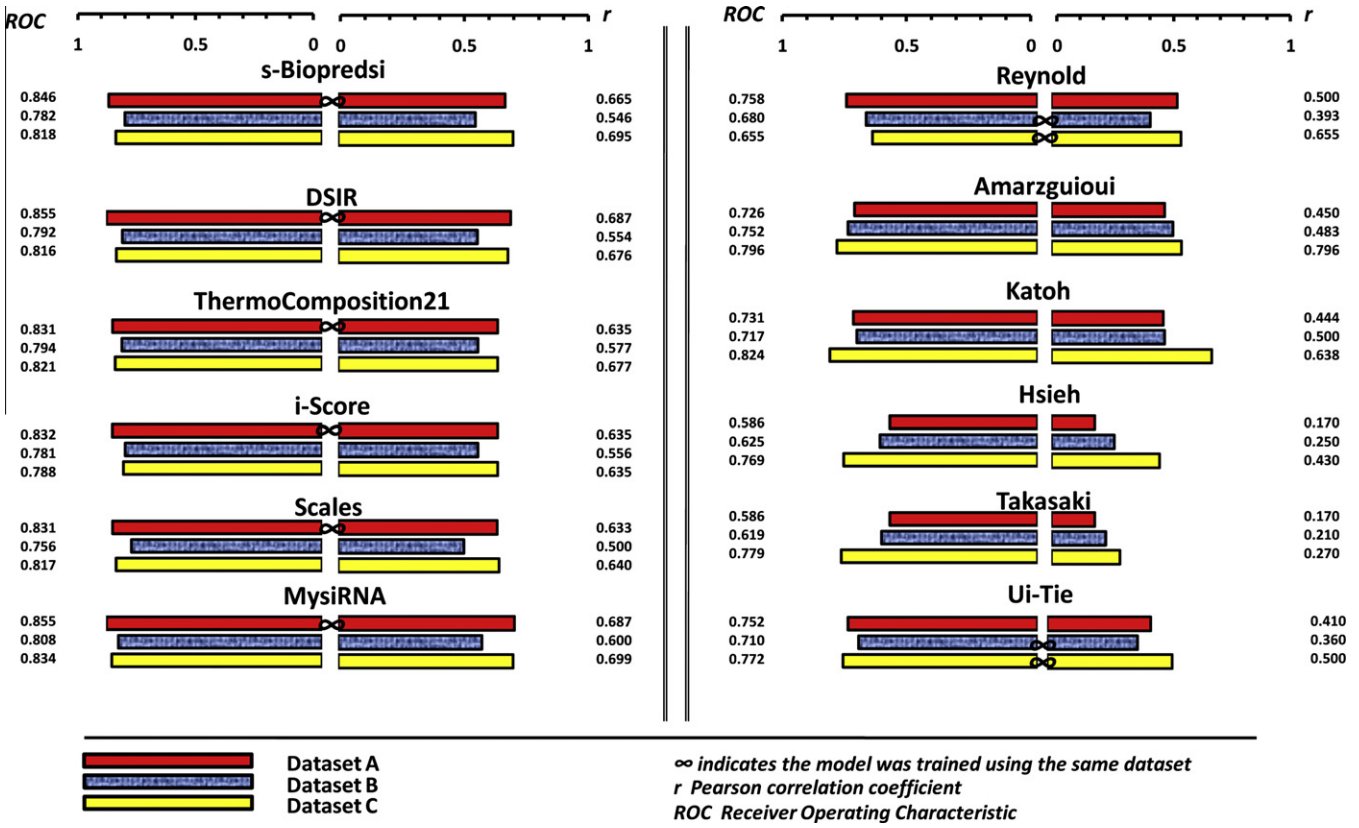

Fig. 4. Comparative studies between MysiRNA, first and second generation tools. Comparison between 12 different tools using ROC & Pearson correlation coefficient against dataset A, B and C. The enhancement brought by the second generation tool proven from the finding of this analysis. Additionally, MysiRNA was found to perform with high level of specificity and sensitivity compared to the other models. In cases the mark (∞) indicates possible result overestimation, due to the use of concerned dataset in the training of the model.

Table 3
Comparative study between *MysiRNA Model* and each of the second generation tools. The results illustrate the common true positive and common true negative between MysiRNA and each of the second generation models. Further analysis for was performed for instances showing the enhancement brought by MysiRNA in numerical and percentage against dataset B which has not been used to any model training.

2nd Generation tool	Number of TP and TN in common	Number of FP and FN in common	Number of TP/TN by MysiRNA and FN/FP by 2nd generation tool	Number of FP/FN by MysiRNA and TN/TP by 2nd generation tool	% Improvement brought by MysiRNA (%)
s-Biopredsi	249	79	58	33	6
DSIR	278	94	29	18	3
ThermoComposition21	279	96	28	16	3
i-Score	165	44	142	68	18
Scales	264	94	43	18	6

Table 4
Mathew correlation coefficient comparative analysis. Data of 18,593 recently presented in [25] was used in a comparative analysis involving Ui-Tei, Amarzguioui, Hsieh, Takasaki, Biopredsi, i-Score, Reynolds, Katoh, DSIR and ThermoComposition21, siRNA Scales and MysiRNA models. All the positive records were isolated and combined with negative records of the same size, to test these models. Mathew correlation coefficient (MCC) was used to illustrate the statistical differences between the models involved.

476 Records	*Ui-Tei*	*Amar*	*Hsieh*	*Taka*	*Biopred*	*i-Score*	*Rey*	*Katoh*	*DSIR*	*Thermo21*	*Scales*	*MysiRNA*
Senstitive	0.26	0.06	0.00	0.02	0.18	0	0.067	0.3	0.44	0.4	0.34	0.64
Specificity	0.95	0.99	0.99	0.99	0.98	1	0.98	0.89	0.94	0.91	0.94	0.86
mmc	0.29	0.12	0	0.05	0.26	–	0.13	0.23	0.43	0.37	0.35	0.51
TP	61	14	1	6	43	0	16	72	105	96	81	153
FN	177	224	237	232	195	238	222	166	133	142	157	85
TN	226	235	237	235	233	238	234	211	223	217	224	204
FP	12	3	1	3	5	0	4	27	15	21	14	34

Three validation experiments were conducted to test our proposed model using Pearson correlation coefficient, ROC analysis and MCC for validation. In these comparative studies, it was found that our *MysiRNA* model is able to predict siRNA inhibition efficiency with better specificity and selectivity than any previous model.

5. Conclusions

This work introduces *MysiRNA*, featuring an artificial neural network model, to predict siRNA inhibition activity, built on two previous models (*ThermalComposition21* and *i-Score*) and whole stacking energy (ΔG). It may be considered as an example of successfully combining different machine learning methods (classifiers) to improve the prediction accuracy compared to previously designed models. This illustrates the improvement in performance that can sometimes be achieved by incorporating different models in a non-linear model combination. As has previously been shown in other areas of bioinformatics, the importance of consensus results (combining results between various algorithms) appears to be valid here. MysiRNA score is implemented in our siRNA design software, MysiRNA-Designer [1], where it was found able to boost the specificity from 93% to 97%. *MysiRNA* WEKA Model is freely available in the supplementary data with the Perl script to run it. It is expected that *MysiRNA* may become widely used for siRNA selection. In addition, the approach of combining multiple classifiers used here could help improve the performance of prediction in similar areas of bioinformatics.

Acknowledgments

We thank Kinji Ohno, S. Tarek, A. Ahmed and M. ElHadidi for helpful discussions comments and technical assistance. We appreciate the kind help and support of Information Technology Institute (Egypt), National Research Center (Egypt), the University of Nottingham (UK).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbi.2012.02.005.

References

[1] Mysara M, Garibaldi JM, Elhefnawi M. MysiRNA-Designer a workflow for efficient siRNA design. PLoS One 2011;6(10):e25642.
[2] Ullu E, Djikeng A, Shi H, Tschudi C. RNA interference: advances and questions.. Philos Trans R Soc London. Series B, Biol Sci 2002;357:65–70.
[3] Stark GR, Kerr IM, Williams BR, Silverman RH, Schreiber RD. How cells respond to interferons. Ann Rev Biochem 1998;67:227–64.
[4] Hamilton a J, Baulcombe DC. A species of small antisense RNA in posttranscriptional gene silencing in plants. Science (New York NY) 1999;286:950–2.
[5] Elbashir SM, Lendeckel W, Tuschl T. RNA interference is mediated by 21- and 22-nucleotide RNAs. Genes Develop 2001:188–200.
[6] Hutvágner G, Zamore PD. A microRNA in a multiple-turnover RNAi enzyme complex. Science (New York NY) 2002;297:2056–60.
[7] Surabhi RM, Gaynor RB. RNA interference directed against viral and cellular targets inhibits human immunodeficiency Virus Type 1 replication. J Virol 2002;76:12963–73.
[8] Xia H, Mao Q, Eliason SL, Harper SQ, Martins IH, Orr HT, et al. RNAi suppresses polyglutamine-induced neurodegeneration in a model of spinocerebellar ataxia. Nat Med 2004;10:816–20.
[9] ElHefnawi M et al. Identification of novel conserved functional motifs across most Influenza A viral strains. Virol J 2011;8:44.
[10] ElHefnawi M. et al. The design of optimal therapeutic small interfering RNA molecules targeting diverse strains of influenza A virus. Bioinformatics 2011.
[11] Haasnoot J, Westerhout EM, Berkhout B. RNA interference against viruses: strike and counterstrike. Nat Biotechnol 2007;25(12):1435–43.
[12] Davis ME, Zuckerman JE, Choi CHJ, Seligson D, Tolcher A, Alabi CA, et al. Evidence of RNAi in humans from systemically administered siRNA via targeted nanoparticles. Nature 2010;464:1067–70.
[13] Reynolds A, Leake D, Boese Q, Scaringe S, Marshall WS, Khvorova A. Rational siRNA design for RNA interference. Nat Biotechnol 2004;22:326–30.
[14] Amarzguioui M, Prydz H. An algorithm for selection of functional siRNA sequences. Biochem Biophys Res Commun 2004;316:1050–8.
[15] Takasaki S, Kotani S, Konagaya A. An effective method for selecting siRNA target sequences in mammalian cells. Cell Cycle (Georgetown Tex) 2004;3:790–5.
[16] Katoh T, Suzuki T. Specific residues at every third position of siRNA shape its efficient RNAi activity. Nucleic Acids Res 2007;35:e27.
[17] Ui-Tei K, Naito Y, Takahashi F, Haraguchi T, Ohki-Hamazaki H, Juni A et al. Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference; 2004.
[18] Hsieh AC, Bo R, Manola J, Vazquez F, Bare O, Khvorova A, et al. A library of siRNA duplexes targeting the phosphoinositide 3-kinase pathway: determinants of gene silencing for use in cell-based screens. Nucleic Acids Res 2004;32:893–901.
[19] Ren Y, Gong W, Xu Q, Zheng X. siRecords: an extensive database of mammalian siRNAs with efficacy ratings. Access 2006:1–10.
[20] Huesken D, Lange J, Mickanin C, Weiler J, Asselbergs F, Warner J, et al. Design of a genome-wide siRNA library using an artificial neural network. Nat Biotechnol 2006;23:995–1002.
[21] Vert J-P, Foveau N, Lajaunie C, Vandenbrouck Y. An accurate and interpretable model for siRNA efficacy prediction. BMC Bioinform 2006;7:520.
[22] Shabalina SA, Spiridonov AN, Ogurtsov AY. Computational models with thermodynamic and composition features improve siRNA design. BMC Bioinform 2006;7:65.
[23] Ichihara M, Murakumo Y, Masuda A, Matsuura T, Asai N, Jijiwa M, et al. Thermodynamic instability of siRNA duplex is a prerequisite for dependable prediction of siRNA activities. Nucleic Acids Res 2007:1–10.
[24] Matveeva O, Nechipurenko Y, Rossi L, Moore B, Ogurtsov AY, Atkins JF, et al. Comparison of approaches for rational siRNA design leading to a new efficient and transparent method. Access 2007;35:1–10.
[25] Khvorova A, Reynolds A, Jayasena SD. Functional siRNAs and miRNAs exhibit strand bias. October 2003;115:209–16.
[26] Gong W, Ren Y, Xu Q, Wang Y, Lin D, Zhou H, et al. BMC Bioinform 2006;7:516.
[27] Patzel V. In silico selection of active siRNA. Drug Discov Today 2007;12:139–48.
[28] Ladunga I. More complete gene silencing by fewer siRNAs: transparent optimized design and biophysical signature. Nucleic Acids Res 2007;35:433–40.
[29] Fellmann C, Zuber J, McJunkin K, Chang K, Malone CD, et al. Functional identification of optimized RNAi triggers using a massively parallel sensor assay. Mol cell 2011;41:733–46.